fastMitoCalc Tutorial

Prepared by Yong Qian and Jun Ding

Updated on Dec 2, 2016

fastMitoCalc is an ultra-fast version of our original program *mitoCalc* that estimates mtDNA copy number from whole-genome sequencing data. *fastMitoCalc* is ~100 times faster than *mitoCalc* and can analyze an individual bam file (with 4X average coverage for nuclear DNA) in less than a minute. So *fastMitoCalc* is suitable for large-scale whole-genome sequencing projects.

NOTE: *fastMitoCalc* does need to work on the indexed bam files!

1. Compile BaseCoverage.cpp file to generate the executable “BaseCoverage”:

   ```
   g++ BaseCoverage.cpp -o BaseCoverage
   ```

2. Run the following command with your own specifications to analyze one individual’s sequencing data (i.e., one bam file):

   ```
   perl fastMitoCalc.pl [options] –f <in.bam> -w <workDir> -p
   <BaseCoverage>
   ```

   - **fastMitoCalc.pl** is the perl file you will be running to execute the fastMitoCalc program. If you wish to run the program outside of your fastMitoCalc folder you can also run “perl /Path/To/Command/fastMitoCalc.pl” as the first part of the command.
   - **<in.bam>** is the individual bam file to be analyzed, and it can be in the format of /Path/To/bam/File/ID.bam. "ID" will be used as the identifier for output files.
   - **<workDir>** is the working directory, also formatted as “/Path/To/workDir”. This is a folder you should create before running the program, and is where the intermediate files and final output files will be generated.
   - **<BaseCoverage>** is the path to the executable "BaseCoverage" in the format of “/Path/To/Command/BaseCoverage”.
   - **[options]: fastMitoCalc.pl** provides multiple options to analyze the data; type “perl fastMitoCalc.pl” to see all the available options (some options are included below as examples).
   - **Example commands** with different **options** to run the program:

   1) perl fastMitoCalc.pl **-n 3000 -s 1000** -f <in.bam> -w <workDir> -p <BaseCoverage>

This will estimate the nuclear DNA sequencing coverage using 3,000 randomly selected 1,000-bp regions. (When –n and –s are not specified, this will be the default option the program will run on). This command also assumes that autosomal chromosomes are labeled as 1, 2, ..., 22, and mtDNA is labeled as MT in the provided bam files. See the next example for options to specify labels for chromosomes in bam files.

2) perl fastMitoCalc.pl –e chr –m chrM -f <in.bam> -w <workDir> -p <BaseCoverage>
   This will estimate the nuclear DNA sequencing coverage with the default setting. “-e chr” and “-m chrM” specify that autosomal chromosomes are labeled as chr1, chr2, ..., chr22, and mtDNA is labeled as chrM in the provided bam files.

3) perl fastMitoCalc.pl **-c 22** -f <in.bam> -w <workDir> -p <BaseCoverage>
   This will estimate the nuclear DNA sequencing coverage using whole chromosome 22 only.

4) perl fastMitoCalc.pl **-b <bedfile>** -f <in.bam> -w <workDir> -p <BaseCoverage>
   This will estimate the nuclear DNA sequencing coverage using regions specified by users in the bed format (Format: chr start_position end_position; “chr” should be specified in the same way as in the bam file). For reproducibility purpose, we have included in the package a default bed file “default.bed” that included 3,000 selected regions of 1,000 bps from 22 chromosomes. Note: this bed file is only useful when chromosome is specified in the same way in both the bed and bam files.
   NOTE: In the download package, we also provide a bed file "1000G_Phase3_20130108.exome.offtargets.bed". This file can be used to do a quick analysis using off-target reads from whole-exome data. This file was created based on the consensus "on-target" regions provided by 1000 Genomes Project ["20130108.exome.targets.bed" on their FTP site, which took the union of design files from two platforms (NimbleGen EZ_exome v1 and Agilent sure select v2)]. Specifically, our "off-target" bed file excluded all the regions plus 50kb up- and down-stream in the "20130108.exome.targets.bed" file. Again, this file is for a quick analysis, but ideally an "off-target" bed file should be created based on the specific exome capture kit used in the sequencing experiment. Also, this bed file is only useful when chromosome is specified in the same way in both the bed and bam files.

3. After the analysis is done, intermediate files will be deleted automatically and the output file will be saved in <workDir> with file name "<ID>MTData.txt". Here is a detailed explanation of the output:

mt_copy_number_avg: the estimated mtDNA copy number, calculated using randomly selected regions, a specific chromosome, or specific regions defined by users.

mt_coverage: the average mtDNA coverage

autosomal_coverage: the average autosomal coverage

actual basepairs covered by reads: total number of bases that are covered by at least one sequencing read in the randomly selected regions, in one specific chromosome, or in regions specified by users in a bed file

chrom_used_for_autosomal_coverage: a list of autosomal chromosomes used to calculate nuclear DNA sequencing coverage